# Modeling Fusion Data in Probabilistic Metric Spaces for the Identification of Confinement Regimes and Scaling Laws

G. Verdoolaege and G. Van Oost

Department of Applied Physics, Ghent University, Belgium

*Corresponding Author:* geert.verdoolaege@ugent.be

**Abstract:**
Pattern recognition plays an important role and has great potential in fusion data analysis. However, a drawback is that individual measurements are usually represented as unstructured points in a Euclidean data space. We argue that a fundamentally probabilistic approach offers significant advantages. It allows representing the data in a non-Euclidean probabilistic space, wherein the patterns of interest are much more distinct, simply because they are based on more information. In this work, we address the identification of confinement regimes and the establishment of a scaling law for the energy confinement time, using data from the International Global H-mode Confinement Database. We propose a single-level and a Bayesian multilevel model for capturing the statistical data uncertainty. We then show that pattern recognition operations working in the associated probability space are considerably more powerful than their counterparts in a Euclidean data space. This opens up new possibilities for analyzing confinement data and for fusion data processing in general.

## 1 Introduction

Pattern recognition can be used in nuclear fusion data analysis for uncovering data structures, such as clusters and regression surfaces, that provide insight into the underlying physical processes. This complements physics studies and contributes to real-time plasma control.

Measurements obtained in fusion experiments can be affected by a considerable uncertainty, which is usually regarded as a nuisance for pattern recognition operations. However, the probability distribution associated to data uncertainties is in fact a very useful piece of information. We here advocate the point of view that the fundamental object resulting from the measurement process is a probability distribution. This probability distribution may contain significantly more information than a measurement value and an error bar alone. Moreover, the probabilistic model can be chosen to embody any information regarding the underlying physics generating the data, the experimental conditions under which the data were collected and the uncertainties affecting the measurements and

model parameters. This additional information can and should be exploited in any further processing of the data, such as pattern recognition. In this paper we apply pattern recognition in probability spaces for the identification of confinement regimes (classification) and for the derivation of a confinement time scaling law (regression).

# 2 A geometric-probabilistic pattern recognition framework

We briefly sketch the principles of a new pattern recognition framework that was introduced in [1] and [2]. Basically, we apply pattern recognition methods directly in a probabilistic data space, i.e. a space of probability distributions. To this end, we employ the mathematical framework of *information geometry*, wherein a probability distribution function (PDF) is interpreted as a point on a manifold [3]. The PDF parameters provide a coordinate system on the manifold and the Fisher information plays the role of a metric tensor. This permits calculating the *geodesic distance* (GD) as a natural and theoretically well-motivated similarity measure between probability distributions [1]. In this paper we discuss applications that are based on a univariate Gaussian model $\mathcal{N}(\mu, \sigma)$, parameterized by its mean $\mu$ and standard deviation $\sigma$. A closed-form expression exists for the associated GD, permitting fast and accurate computations [1].

# 3 Classification and regression for confinement data

We apply pattern recognition on probabilistic manifolds to confinement data obtained from the International Tokamak Physics Activity (ITPA) Global H-mode Confinement Database (henceforth referred to as 'the ITPA database'), version DB3 13f [4]. We first briefly introduce the database and then propose two probabilistic models (single-level and multilevel) describing the database entries. We then apply pattern recognition methods on the corresponding probabilistic manifolds by calculating geodesic distances.

## 3.1 ITPA database

The data in the ITPA database have been used extensively for determining scaling laws for the energy confinement time, mainly as a function of a set of eight plasma and engineering variables: plasma current ($I_\mathrm{p}$), vacuum toroidal magnetic field ($B_\mathrm{t}$), total power loss from the plasma ($P_\mathrm{loss}$), central line-averaged electron density ($\bar{n}_\mathrm{e}$), plasma major radius ($R$), inverse aspect ratio ($\epsilon$), effective atomic mass ($M_\mathrm{eff}$) and elongation ($\kappa$). We will refer to these variables by the designation $x_\lambda$, with $\lambda = 1, \ldots, 8$. Furthermore, we will use the notation $x_{\lambda,ijk}$ for the individual measurement *value* of variable $x_\lambda$, corresponding to database entry $j$ obtained at tokamak $k$. The index $i$ in principle allows the possibility of multiple measurements (e.g. a time series) for a single database entry. However, in the present case there is only one measurement value per database entry, so $i$ can take only the single value 1 (the notation including $i$ will be useful in the following, nevertheless).

In our experiments we used essentially the same eight variables $x_\lambda$ to discriminate between, roughly, L- and H-mode plasmas (the minor radius $a$ was chosen instead of $\epsilon$). Specifically, all database entries with a confinement mode labeled as H, HGELM, HSELM, HGELMH, HSELMH and LHLHL were considered to belong to the H-mode class, while discharges labeled with L, OHM and RI were assigned to the non-H-mode class, or L-mode for brevity.

## 3.2 Single-level model

Throughout our analysis we assumed that the relative error estimates in the ITPA database pertain to a pure statistical error from which a standard deviation can be calculated. According to the principle of maximum entropy the underlying probability distribution of every measurement $x_{\lambda,ijk}$ is Gaussian with as its mean $\mu_{\lambda,jk}$ the measurement itself and standard deviation $\sigma_{\lambda,jk}$ the error bar. Hence, the likelihood of the data point $x_{\lambda,ijk}$ is given by:

$$x_{\lambda,ijk} \sim \mathcal{N}(x_{\lambda,ijk}|\mu_{\lambda,jk}, \sigma_{\lambda,jk}), \tag{1}$$

where, trivially, the maximum-likelihood (ML) estimate of $\mu_{\lambda,jk}$ is $x_{\lambda,ijk}$. The ML estimate of $\sigma_{\lambda,jk}$ is obviously zero, since we have only a single measurement per database entry. However, in our treatment we will exploit the prior information that tells us that the standard deviation is actually given by the error bar quoted in the database.

In addition, although our framework can perfectly handle multivariate distributions, we here suppose that all variables are statistically independent, so their joint distribution factorizes. It should be noted that this does not exclude at all a deterministic dependence between the variables.

## 3.3 Bayesian multilevel model

Apart from prior information on the standard deviations, there is also useful information available on the means due to additional structure in the database. Indeed, at the minimum we know for each measurement at which tokamak it was obtained. This yields a priori information on the typical range of a variable that is to be expected at a certain machine, just by studying the statistics of the measurements in the database from that machine. Likewise, we can determine statistics concerning the entire database. The multilevel structure of the database can be ideally captured by a *hierarchical model*, specifically a multilevel model, wherein at several levels the data are assumed to be distributed following a specific probabilistic model. Combining these submodels into a single probabilistic model via the standard rules of probability theory, one can subsequently estimate the model parameters via traditional frequentist techniques or Bayesian methods. We here adopt the latter approach because it agrees with our point of view that all measured and estimated quantities are fundamentally uncertain (see [5] for an overview of Bayesian methods and [6] for an accessible account in the context of fusion data analysis).

The first level of our model has already been specified in (1), governing the distribution of the measurements. At the second level, we exploit prior information about the

distribution of the means $\mu_{\lambda,jk}$. Specifically, we put a prior distribution on the parameters $\mu_{\lambda,jk}$, which, in turn, is characterized by a set of *hyperparameters*. Indeed, for fixed $k$ we may consider each $\mu_{\lambda,jk}$ as a value drawn from a distribution characterizing the variability of $\mu_{\lambda,jk}$ at tokamak $k$. Thus, the variability is considered fully 'stochastic' (or 'random') by nature, although it is more accurate to state that, at this point, we do not have—or we choose to neglect—further information that could be used to describe the data-generating mechanism in more detail.

In addition, it is important to note that in a fully Bayesian treatment, no data is to be used for constructing the prior distribution or for estimating its parameters, in order to avoid circular arguments. Nevertheless, although in the same spirit one should also consider a prior distribution on the standard deviation parameter $\sigma_{\lambda,jk}$, in this work we make an approximation by first estimating all parameters representing a standard deviation and keeping them fixed throughout the analysis. From the Bayesian viewpoint, this means that in our analysis standard deviations are not considered as parameters, but are constants.

Various principles exist that assist in the selection of a prior distribution, but in a multilevel analysis the concept of *conjugacy* of prior distributions with respect to the likelihood is a popular criterium. The advantage is that the posterior distribution is then part of the same family of probability distributions as the prior, so the mathematics remain tractable. The conjugate distribution of the normal distribution with known standard deviation is again normal. Hence, each $\mu_{\lambda,jk}$ is assumed to be distributed according to a normal distribution with a certain mean $\mu_{\lambda,k}$ and standard deviation $\sigma_{\lambda,k}$. Again, we make an approximation by taking $\sigma_{\lambda,k}$ constant for a given $\lambda$ and $k$, namely the sample standard deviation of the set of measurements $x_{\lambda,ijk} = \mu_{\lambda,jk}$, for all $j$ at tokamak $k$, which is in fact the ML estimate of $\sigma_{\lambda,k}$.

Similarly, we model a third level, where each $\mu_{\lambda,k}$ is normally distributed with mean $\mu_{\lambda,0}$ and standard deviation $\sigma_{\lambda,0}$, the latter estimated by the sample standard deviation of all measurements of the variable $x_\lambda$ over the complete database.

At the fourth level we assume a normal distribution for $\mu_{\lambda,0}$, characterized by a mean parameter $\phi_\lambda$ and standard deviation $\tau_\lambda$. A fifth and final level cuts off the hierarchy by assuming that $\phi_\lambda$ is uniformly distributed, while $\tau_\lambda$ is defined as a fixed percentage of $\phi_\lambda$ (here we chose 10%). We can be vague about these parameters since they are high up the hierarchy and it can be proved that a uniform prior leads to a proper posterior distribution in a Gaussian hierarchy [5]. The complete hierarchical model has been summarized in Table I.

The joint posterior distribution related to the variable $x_\lambda$ for the complete set of parameters that are to be estimated is then given by the following proportionality relation:

$$p(\mu_{\lambda,jk}, \mu_{\lambda,k}, \mu_{\lambda,0}, \phi_\lambda | x_{\lambda,ijk}, \forall j, k) \sim \prod_{j,k} \mathcal{N}(x_{\lambda,ijk}|\mu_{\lambda,jk}, \sigma_{\lambda,jk}) \mathcal{N}(\mu_{\lambda,jk}|\mu_{\lambda,k}, \sigma_{\lambda,k}) \mathcal{N}(\mu_{\lambda,k}|\mu_{\lambda,0}, \sigma_{\lambda,0}) \mathcal{N}(\mu_{\lambda,0}|\phi_\lambda, \tau_\lambda). \quad (2)$$

Among the various ways to simulate (sample) from this distribution, we chose a simple Gibbs sampling procedure, first establishing the posterior distribution for each individual

TABLE I: A MULTILEVEL BAYESIAN STRUCTURE USED FOR MODELING THE ITPA DATA FOR VARIABLE $x_\lambda$ ($\lambda = 1, \ldots, 8$).

| Level | Model |
|---|---|
| 1 | $x_{\lambda,ijk} \sim \mathcal{N}(x_{\lambda,ijk}\|\mu_{\lambda,jk}, \sigma_{\lambda,jk})$ |
| 2 | $\mu_{\lambda,jk} \sim \mathcal{N}(\mu_{\lambda,jk}\|\mu_{\lambda,k}, \sigma_{\lambda,k})$ |
| 3 | $\mu_{\lambda,k} \sim \mathcal{N}(\mu_{\lambda,k}\|\mu_{\lambda,0}, \sigma_{\lambda,0})$ |
| 4 | $\mu_{\lambda,0} \sim \mathcal{N}(\mu_{\lambda,0}\|\phi_\lambda, \tau_\lambda)$ |
| 5 | $\phi_\lambda \sim \mathcal{U}(-\infty, +\infty), \quad \tau_\lambda = 0.1\phi_\lambda$ |

parameter, conditional on all others [5]. Summarizing the complete data set corresponding to variable $x_\lambda$ in the data vector $\boldsymbol{x}_\lambda$, this leads for the normal model to simple and intuitive expressions. Each parameter is fully determined by the parameters with the same indices $j$ and $k$ at the level below, as well as by the constant standard deviations. Hence, we have:

$$\mu_{\lambda,jk}|\mu_{\lambda,k}, \mu_{\lambda,0}, \phi_\lambda, \boldsymbol{x}_\lambda \sim \mathcal{N}(\hat{\mu}_{\lambda,jk}, \hat{\sigma}_{\lambda,jk}), \quad \hat{\mu}_{\lambda,jk} = \frac{\frac{\mu_{\lambda,k}}{\sigma^2_{\lambda,k}} + \frac{n_{jk}}{\sigma^2_{\lambda,jk}}\bar{x}_{\lambda,\cdot jk}}{\frac{1}{\sigma^2_{\lambda,k}} + \frac{n_{jk}}{\sigma^2_{\lambda,jk}}}, \quad \hat{\sigma}^2_{\lambda,jk} = \frac{1}{\frac{1}{\sigma^2_{\lambda,k}} + \frac{n_{jk}}{\sigma^2_{\lambda,jk}}},$$

$$\mu_{\lambda,k}|\mu_{\lambda,jk}, \mu_{\lambda,0}, \phi_\lambda, \boldsymbol{x}_\lambda \sim \mathcal{N}(\hat{\mu}_{\lambda,k}, \hat{\sigma}_{\lambda,k}), \quad \hat{\mu}_{\lambda,k} = \frac{\frac{\mu_{\lambda,0}}{\sigma^2_{\lambda,0}} + \frac{n_k}{\sigma^2_{\lambda,k}}\bar{\mu}_{\lambda,\cdot k}}{\frac{1}{\sigma^2_{\lambda,0}} + \frac{n_k}{\sigma^2_{\lambda,k}}}, \quad \hat{\sigma}^2_{\lambda,k} = \frac{1}{\frac{1}{\sigma^2_{\lambda,0}} + \frac{n_k}{\sigma^2_{\lambda,k}}},$$

$$\mu_{\lambda,0}|\mu_{\lambda,jk}, \mu_{\lambda,k}, \phi_\lambda, \boldsymbol{x}_\lambda \sim \mathcal{N}(\hat{\mu}_{\lambda,0}, \hat{\sigma}_{\lambda,0}), \quad \hat{\mu}_{\lambda,0} = \frac{\frac{\phi_\lambda}{\tau^2_\lambda} + \frac{n_m}{\sigma^2_{\lambda,0}}\bar{\mu}_{\lambda,\cdot}}{\frac{1}{\tau^2_\lambda} + \frac{n_m}{\sigma^2_{\lambda,0}}}, \quad \hat{\sigma}^2_{\lambda,0} = \frac{1}{\frac{1}{\tau^2_\lambda} + \frac{n_m}{\sigma^2_{\lambda,0}}},$$

$$\phi_\lambda|\mu_{\lambda,jk}, \mu_{\lambda,k}, \mu_{\lambda,0}, \boldsymbol{x}_\lambda \sim \mathcal{N}(\hat{\phi}_\lambda, \hat{\tau}_\lambda), \quad \hat{\phi}_\lambda = \mu_{\lambda,0}, \quad \hat{\tau}^2_\lambda = \tau^2_\lambda.$$

Here, the notation of an index that is replaced by a dot, accompanying a quantity with an overbar, refers to an averaging process over all measurements or parameters in the range of that index. For example, $\bar{x}_{\lambda,\cdot jk}$ is just the measurement $x_{\lambda,1jk}$ and $\bar{\mu}_{\lambda,\cdot k}$ is the parameter obtained by taking the average of all parameters $\mu_{\lambda,jk}$, running over all $j$ for a fixed $k$. In addition, $n_{jk}$ denotes the number of samples per database entry (only one for every variable), $n_k$ is the number of database entries for machine $k$ and $n_m$ is the number of machines in the database (i.e. 19). According to the expressions above, the posterior mean parameter at each level is given by an average of the information provided by the data and by the prior distribution, weighted by the respective inverse variances.

Given the Gaussian posterior conditionals, the Gibbs sampling procedure is very simple. Starting from some initial values of the parameters (e.g. the ML estimates), one samples from the posterior conditionals a new value for each of the parameters in series, given the current values of the other parameters. Carrying out this process iteratively, it can be proved that, after convergence has been reached, this algorithm generates samples from each of the marginal posterior distributions (and from the joint posterior as well) [5]. As soon as a sufficiently large sample size has been reached, one can estimate the posterior

TABLE II: CORRECT CLASSIFICATION RATES (%) OF CONFINEMENT REGIMES USING A KNN CLASSIFIER FOR EUCLIDEAN AND GEODESIC DISTANCE MEASURES.

| Mode | Euclidean w/o errors | Euclidean with errors | GD single-level | GD multilevel |
|---|---|---|---|---|
| L | 89.7 | 91.2 | 91.9 | 93.2 |
| H | 89.1 | 90.5 | 93.3 | 94.3 |

mode (leading to maximum-a-posteriori parameter estimates) moments (mean parameter estimates, variance), credible intervals (the Bayesian analog of confidence intervals), etc.

In the confinement mode classification problem described below, we will calculate GDs on the combined probabilistic manifold corresponding to the posterior distribution parameterized by the $\mu_{\lambda,jk}$, $\mu_{\lambda,k}$, $\mu_{\lambda,0}$ and $\phi_\lambda$, given in (2). For each database entry we use the posterior mean of each of the parameters, estimated from the Gibbs sampling procedure, together with the assigned standard deviations, as coordinates on the manifold.

## 3.4 Confinement mode classification

We now apply a classification algorithm for the automated identification of confinement regimes, basically L-mode and H-mode. This has important applications in plasma control and will be an powerful tool for ITER. Another application to disruption prediction was described in [1].

We performed a series of classification experiments with two classes (L- and H-mode) using 1% of the data for training. We carried out $k$-nearest neighbor (kNN) classification with $k = 1$, effectively assigning a point to be classified to the class that its nearest neighbor belongs to. The experiments were performed once without and once with consideration of the measurement error. In the latter case, we applied both the single-level and multilevel model introduced above.

The results are shown in Table II. The correct classification rates for both the L-mode and H-mode data are clearly better if the measurement error is considered, even using the Euclidean distance. The GD performs better than the Euclidean distance, since the GD properly takes into account the geometry of the probabilistic manifold. Finally, the multilevel model is seen to provide an advantage compared to the single-level model.

## 3.5 Confinement time scaling law

We next consider regression in the ITPA database, with the purpose of obtaining a scaling law for the global energy confinement time $\tau_{\rm E}$, as a function of the eight variables introduced in Section 3.1. Our method is based on geodesic regression (GR), which was described in [2]. Basically, instead of minimizing the sum of Euclidean distances between the data points and the values predicted by the regression model, in GR the sum of

TABLE III: COEFFICIENT OF DETERMINATION $R^2$ FOR SEVERAL REGRESSION METHODS.

| OLS | TLS | GR single | GR multi | OLS in GR single | TLS in GR single |
|---|---|---|---|---|---|
| 0.94 | 0.97 | 0.71 | 0.78 | 0.44 | 0.52 |

GDs is minimized between the data probability distributions and the predicted distributions assuming a linear regression model. At this point we do not attempt to verify existing scaling laws for $\tau_{\mathrm{E}}$. Rather, we aim to show that our method, which minimizes geodesic distances, yields an enhanced goodness-of-fit of the regression function on the probabilistic manifold, compared to regular regression based on the minimization of Euclidean distances. To do this, we calculated the well-known coefficient of determination $R^2$. In order to evaluate GR, we adapted the definition for $R^2$ by using GDs and geodesic centroids [2].

We applied several methods in a regression analysis for $\tau_{\mathrm{E}}$ on the so-called 'standard' set of the ITPA data. The methods with which we compare geodesic regression are ordinary least squares (OLS) and total least squares (TLS) (errors in variables) using singular value decomposition, which allows errors on the independent variables as well. The GR was carried out once with the single-level data model and once with the multilevel model.

The results of this study are summarized in Table III, mentioning the coefficients of determination $R^2$ resulting from the various regression methods. It should be noted that the $R^2$ values of GR cannot simply be compared to those found with OLS and TLS, since different objects are fitted with the respective regression functions (structureless points vs. PDFs). However, we can demonstrate that GR better takes into account the intrinsic uncertainty on each measurement, by substituting the values of the coefficients resulting from OLS and TLS into the calculation of $R^2$ based on the GD with the single-level model. The resulting values are reported in Table III as well, and are labeled by 'OLS in GR single' and 'TLS in GR single'. The $R^2$ value obtained with GR and the single-level model ('GR single') is indeed significantly higher compared to the 'OLS in GR single' and 'TLS in GR single' values. The best fit, however, is provided by GR in conjunction with the multilevel model, proving again that the additional database structure embodied by this model is very relevant to the regression problem.

# 4 Conclusions and outlook

We have argued that a fundamentally probabilistic modeling of the data is required and beneficial in pattern recognition applications for fusion data. This approach is very different from the traditional modeling of data as structureless points in a Euclidean space, in that the probabilistic structure (including the error bars) of the data *actively* helps determining the data patterns (clusters, regression functions, etc.), from the very start of the analysis.

We wish to stress that in these experiments very little information was used regarding the structure of the database, physical interpretation of the data, experimental conditions at each machine and the data probability distribution. The performance differences between the various methods are relatively small, yet it is remarkable that such little differences in knowledge states can make a noticeable difference in classification rates. While our two simple models (single-level and multilevel) capture only the broad outlines of the data structure, there is an enormous amount of additional information that one can introduce into the probabilistic model describing the data. For instance, one may encode prior information regarding the experimental conditions during the discharges from which the ITPA database entries were sampled, include predictions and trends from existing theoretical models, etc. Especially in the case of the regression problem for scaling laws, such an approach could make a substantial difference compared to some of the traditional approaches, where the data are simply regarded as a 'cloud' in the data space, without much memory of the underlying physics that generated the data in the first place. The additional information is available and extremely relevant, and the best way of making use of it is by modeling the data with a probability model that includes this information. Future work will mainly explore this exciting possibility, with further applications to various pattern recognition problems in fusion science.

# References

[1] G. Verdoolaege, G. Karagounis, A. Murari, J. Vega, G. Van Oost, and JET-EFDA Contributors. Modeling fusion data in probabilistic metric spaces: Applications to the identification of confinement regimes and plasma disruptions. *Fusion Sci. Technol.*, 62(2):356–365, 2012.

[2] G. Verdoolaege, G. Karagounis, M. Tendler, and G. Van Oost. Pattern recognition in probability spaces for visualization and identification of plasma confinement regimes and confinement time scaling. *Plasma Physics and Controlled Fusion, in press*, 2012.

[3] S. Amari and H. Nagaoka. *Methods of information geometry*, volume 191 of *Transactions of mathematical monographs*. American Mathematical Society, New York, 2000.

[4] D.C. McDonald et al. Recent progress on the development and analysis of the ITPA global H-mode confinement database. *Nucl. Fusion*, 47(3):147–174, 2007.

[5] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian data analysis*. Chapman & Hall/CRC, Boca Raton, second edition, 2004.

[6] G. Verdoolaege, R. Fischer, G. Van Oost, and JET-EFDA Contributors. Potential of a Bayesian integrated determination of the ion effective charge via bremsstrahlung and charge exchange spectroscopy in tokamak plasmas. *IEEE Trans. Plasma Sci.*, 38(11):3168–3196, 2010.